# Increasing the precision of orthology-based complex prediction through network alignment

Macromolecular assemblies play an important role in almost all cellular processes. However, despite several large-scale studies, our current knowledge about protein complexes is still quite limited, thus advocating the use of *in silico* predictions to gather information on complex composition in model organisms. Since protein-protein interactions present certain constraints on the functional divergence of macromolecular assemblies during evolution, it is possible to predict complexes based on orthology data. Here, we show that incorporating interaction information through network alignment significantly increases the precision of orthology-based complex prediction. Moreover, we performed a large-scale *in silico* screen for protein complexes in human, yeast and fly, through the alignment of hundreds of known complexes to whole organism interactomes. Systematic comparison of the resulting network alignments to all complexes currently known in those species revealed many conserved complexes, as well as several novel complex components. In addition to validating our predictions using orthogonal data, we were able to assign specific functional roles to the predicted complexes. In several cases, the incorporation of interaction data through network alignment allowed to distinguish real complex components from other orthologous proteins. Our analyses indicate that current knowledge of yeast protein complexes exceeds that in other organisms and that predicting complexes in fly based on human and yeast data is complementary rather than redundant. Lastly, assessing the conservation of protein complexes of the human pathogen *Mycoplasma pneumoniae*, we discovered that its complexes repertoire is different from that of eukaryotes, suggesting new points of therapeutic intervention, whereas targeting the pathogen's Restriction enzyme complex might lead to adverse effects due to its similarity to ATP-dependent metalloproteases in the human host.

1  List of authors:

2  Roland A. Pache[1,2,*] and Patrick Aloy[1,3]

3  1. Joint IRB-BSC Program in Computational Biology, Institute for Research in

4  Biomedicine (IRB Barcelona), c/ Baldiri Reixac 10-12, 08028 Barcelona, Spain

5  2. Current affiliation: Department of Bioengineering and Therapeutic Sciences,

6  University of California San Francisco (UCSF), 1700 4$^{th}$ street, 94158 San Francisco,

7  California, USA

8  3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys

9  23, 08010 Barcelona, Spain

10  *Corresponding author:

11  Roland A. Pache. Tel: +1 415 514 4307; Email: roland.pache@ucsf.edu

12  Email addresses:

13  RAP: roland.pache@ucsf.edu

14  PA: patrick.aloy@irbbarcelona.org

## Introduction

Almost every major process in a cell, such as replication, transcription, translation and degradation, is carried out not by single proteins, but by macromolecular complexes, regulated through intricate networks of protein-protein interactions. To understand cellular behaviour on a systemic level, we thus need a comprehensive knowledge of the protein complexes present in the respective organism.

In the last years, many small-scale studies have identified protein complexes in yeast and human, which have been collected in the public databases MPACT [1] and CORUM [2], respectively. Moreover, several large-scale proteomics experiments, using tandem-affinity purification coupled to mass spectrometry analysis, have focused on systematically unveiling the composition of macromolecular complexes in the budding yeast *Saccharomyces cerevisiae* [3,4,5,6,7], the human pathogen *Mycoplasma pneumoniae* [8], and recently also in the fruit fly *Drosophila* melanogaster [9] and in human [10,11]. Although yeast is the least complex eukaryotic model organism with about 6,000 genes, several hundreds of protein complexes were found, and the total number of complexes in yeast was estimated to be over 800 [5]. Moreover, the first screen in *M. pneumonia* already yielded 62 homo- and 116 heteromultimeric protein complexes, despite the pathogen's small genome of only 689 protein-coding genes [8]. The importance of protein complexes for cell survival becomes apparent when probing the essentiality of their protein-coding genes through knock-out mutations. Indeed, several studies have shown that protein complexes in yeast are significantly enriched in essential genes [12,13,14,15]. To discover the molecular details of how individual proteins function together as macromolecular assemblies, follow-up initiatives have then aimed at identifying those complexes that are suitable for structural studies by combining systematic bioinformatics and experimental validation strategies [16,17].

So far, however, these important investigations, which have improved our understanding of the architecture and function of protein complexes, were limited to yeast, not only due to the scarcity of functional data in other organisms, but also because their protein complexes are yet largely unknown. To determine whether the findings for yeast complexes describe general principles of molecular organization and to discover how protein complexes have evolved, it is thus necessary to define protein complexes in other species, complementing the results of recent screening efforts [9,10,11]. Since the

47    experimental characterization of macromolecular assemblies is difficult and requires

48    large amounts of time and resources, predicting protein complexes based on existing

49    protein-protein interaction (PPI) and orthology data becomes an interesting alternative.

50    Indeed, different strategies have been developed to exploit these data. On the one hand,

51    several graph-clustering strategies have been applied to interactome networks in order

52    to identify functional modules and protein complexes as densely connected subgraphs

53    [18,19,20]. In addition, new algorithms have specifically been designed for this task with

54    the aim to distinguish real complex components from spurious interactors and to allow

55    shared components across different complexes [21,22]. Various clustering techniques

56    were also used to define protein complexes from purification data in the original large-

57    scale screens of macromolecular assemblies in yeast, fly, human and *M. pneumoniae*

58    [5,6,7,8,10,11]. On the other hand, since protein complexes are often conserved due to

59    the constraints PPIs pose on functional divergence during evolution [23,24], it is possible

60    to predict complexes using orthology information. In its simplest form, orthology-based

61    complex prediction reports the collection of all orthologous proteins of a given complex

62    in one species as the corresponding complex in the other organism [25]. However, one-

63    to-many and many-to-many orthology relationships between species often imply

64    functional divergence of paralogous genes after duplication, leading to the prediction of

65    many false complex components with increasing proteome size.

66    The recent advent of tools for the comparison and alignment of protein interaction

67    networks [26,27,28,29,30,31,32] now opens up new possibilities for complex prediction.

68    One strategy is to align whole interactome networks of different species to search for

69    conserved functional modules [27,32,33,34]. For instance, Sharan *et al.* aligned the

70    yeast and *H. pylori* interactomes, finding 11 conserved protein complexes [27], while

71    Hirsh and colleagues found 150 conserved complexes by aligning the yeast and fly

72    interactomes, matching known complexes in yeast with coherent functional annotations

73    [33]. However, interactome to interactome alignment does not exploit knowledge about

74    the composition of known complexes. This can only be done through complex to

75    interactome alignment, in which the network representation of a known query complex in

76    a given organism such as yeast is aligned to the interactome of a target species. For

77    instance, Dost *et al.* developed the QNet algorithm [35], which allows the querying of

78    input graphs of treelike topology in interaction networks, and used it to align 94 manually

79  curated yeast complexes from the MPACT database [1] to the fly interactome, finding 36

80  of them to be conserved in fly.

81  For this work, we applied our recently developed tool for network comparison,

82  NetAligner, which was demonstrated to outperform the current standard in the field in a

83  variety of different benchmarks [36] and was already used successfully to discover the

84  role of structural disorder in the rewiring of interactomes during evolution [37]. Using

85  NetAligner, we investigated how the incorporation of interaction data through network

86  alignment influences the performance of orthology-based complex prediction. Moreover,

87  we systematically aligned known protein complexes in *H. sapiens*, *S. cerevisiae* and *M.*

88  *pneumoniae* to whole species interactomes to find novel complex components in yeast

89  and human, predict yet undiscovered complexes in the fly *D. melanogaster* and search

90  for similarities and differences between the complexes repertoires of *H. sapiens* and the

91  human pathogen *M. pneumoniae*.

## Results and Discussion

92

### *Network alignment increases precision in orthology-based complex prediction*

93

94  A standard, straightforward method to predict protein complexes in a target species

95  based on those known in a given query organism is the so-called 'orthologs approach'.

96  In that method, the union of all orthologous proteins of the respective query complex

97  components is predicted to constitute that complex in the target species [25]. The

98  general idea behind this approach is that many protein complexes are evolutionarily

99  conserved, because they perform critical cellular tasks, such as replication, transcription

100  or translation, needed in all forms of cellular life. However, due to evolutionary

101  divergence of proteins after duplication, which can lead to functionally non-overlapping

102  paralogs, not all orthologs of the components of a given complex in one species should

103  be expected to be part of the corresponding complex in another organism. Prediction of

104  protein complexes using the standard orthologs approach can thus result in false

105  complex components (i.e. false positive predictions). To test whether the incorporation of

106  interaction information through network alignment can decrease the number of false

107  positives and thus increase the precision of orthology-based complex prediction, we

108    compared the performance of the orthologs approach to that of NetAligner [36] in

109    predicting yeast protein complexes based on human complexes and vice versa through

110    complex to interactome alignment (Fig. 1A). For this, we used the non-redundant

111    benchmark set of 71 matching human-yeast complex pairs (see *Materials & Methods*),

112    which we previously defined [36], and analysed the complex predictions of the orthologs

113    approach with respect to how well they agree with the corresponding benchmark set

114    complexes (see *Materials & Methods*). We evaluated both precision (i.e. fraction of true

115    complex components among all proteins predicted to be part of the given complex) and

116    recall (i.e. fraction of complex components recovered in the given prediction) in the

117    same way as in our previous study [36], so that we could directly compare the

118    performance of the orthologs approach to the performance reported for NetAligner in

119    predicting protein complexes via complex to interactome alignment (see below and

120    *Materials & Methods*). We found that incorporating interaction data through network

121    alignment significantly increases the precision of orthology-based complex prediction

122    (i.e. the ability to distinguish orthologs that are part of the complex in the target species

123    from those that are not). When using NetAligner with default parameters there is a

124    significant increase in precision from 34.6% to 54.1% (p-value=$2.04{\times}10^{-32}$, one-sided

125    Fisher's exact test), which is also present when calibrating NetAligner for the alignment

126    of human complexes to the yeast interactome (53.9%, p-value=$4.07{\times}10^{-32}$, one-sided

127    Fisher's exact test) or vice versa (58.5%, p-value=$2.65{\times}10^{-36}$, one-sided Fisher's exact

128    test; Fig. 1B). This increase in precision arises from NetAligner using protein-protein

129    interaction data to identify and filter out false positive orthology mappings. Due to current

130    interactome networks still being incomplete and missing many interactions existing in

131    nature [38], this filtering step unavoidably leads to an increase in false negatives and

132    thus a reduced recall of true complex components (from 58.0% to 44.9% with default

133    parameters; Fig. 1B). This is caused by orthologous proteins getting filtered out of

134    complex predictions, since they seem not to interact with any other complex component,

135    while those interactions indeed do exist and are just missing from current interactome

136    networks. So currently, there is a tradeoff between precision and recall when using

137    NetAligner to predict protein complexes. However, once species interactomes become

138    more complete, we would expect network alignment approaches to deliver increased

139    precision at the same level of recall of simple orthology mapping. Ongoing and future

140 interaction detection experiments should increase interactome coverage and thus also
141 increase the recall of true complex components when predicting protein complexes
142 through network alignment.

143 ***Strategy to predict complexes through complex to interactome alignment***
144 To predict complexes in yeast, human and fly, which are the species with the most
145 interaction data available, and to identify novel complex components, we aligned the
146 non-redundant sets of human, yeast and mycoplasma protein complexes (see *Materials*
147 *& Methods*) to whole species interactomes using NetAligner [36]. In contrast to a pure
148 orthologs-approach, network alignment via NetAligner incorporates knowledge about
149 protein-protein interactions into orthology-based complex prediction. When aligning a
150 query complex to the interactome of a target species, NetAligner aligns those pairs of
151 orthologous proteins that are part of the two input networks and identifies conserved and
152 likely conserved interactions, as well as parts where the query complex and target
153 interactome differ slightly, represented through gaps and mismatches in the alignment
154 graph constructed by the program [36]. Several program parameters, such as the vertex
155 and edge probability thresholds, further determine which pairs of orthologous proteins
156 will be part of the final alignment solutions. For those alignment scenarios that were part
157 of the complex to interactome alignment benchmark reported in [36] (i.e. yeast to human
158 and human to yeast), we could use the best-performing parameter combinations, while
159 for the others there is no benchmark data available, and we thus used the default
160 parameters (Table 1). We considered only the highest-ranked significant alignment
161 solution (with a standard p-value threshold of 0.05) for each query complex.

162 ***Finding novel components of yeast and human protein complexes***
163 To find novel components of protein complexes in yeast and human, we aligned the non-
164 redundant sets of 1027 human complexes and 244 yeast complexes from the manually
165 curated databases CORUM [2] and MPACT [1], to the yeast and human interactome,
166 respectively (see *Materials & Methods*). Here, we did not use the complexes identified in
167 large-scale studies to ensure that our predictions are based only on curated data
168 sources. This yielded 257 non-redundant significant complex predictions in yeast and 89
169 in human (Table 2 and Fig. 2A & 2C). We then identified novel components by
170 comparing our complex predictions to all known complexes in the respective species

171 (see *Materials & Methods*). We found 307 non-redundant novel components across 181

172 yeast complexes and 175 non-redundant novel components in 65 human complexes

173 (Table 2 and Fig. 2B & 2D). Given the recall of the method when aligning human

174 complexes to the yeast interactome and vice versa (Table 1), most of the predicted

175 complexes are probably sub-complexes (as our method misses some complex

176 components due to incomplete interaction data). Moreover, based on the precision of

177 our method (Table 1), we can estimate that at least 60% of the novel yeast complex

178 components we predicted (from aligning human complexes to the yeast interactome)

179 and 49% of the novel human components we predicted (from aligning known yeast

180 complexes to the human interactome) are real complex members. For an independent

181 *in silico* validation of our predictions, we computed the number of predicted complexes

182 that are functionally homogeneous (see *Materials & Methods*) separately for each Gene

183 Ontology (GO) category (i.e. biological process, molecular function and cellular

184 component) [39], and compared it to the respective number of functionally

185 homogeneous query complexes and predictions based on simple orthology mapping

186 (Fig. 3A & 3B). We found that, even when requiring the complexes to fulfill this criterion

187 in at least two GO categories (e.g. 'biological process' and 'molecular function'; see

188 *Materials & Methods*), the majority of the predicted complexes in yeast (70%) and

189 human (73%) are indeed functionally homogeneous. The fact that the fraction of human

190 query complexes that are homogeneous is smaller than the fraction of predicted yeast

191 complexes (63% vs. 70%) might indicate that our predictions are, on average, of similar

192 quality as the manually-curated human complexes stored in the CORUM database [2].

193 When using a simple orthologs approach instead of NetAligner, we got a similar fraction

194 of homogeneous complex predictions in yeast (73%). On the other hand, the fraction of

195 homogeneous yeast query complexes is larger than the fraction of homogeneous

196 complex predictions in human (97% vs. 73%), suggesting that in this case, our

197 predictions are of lower quality than the query complexes. However, when using a

198 simple orthologs approach, the resulting fraction of homogeneous complex predictions

199 in human is even lower with only 58% (Fig. 3A & 3B). A likely explanation for these

200 results is that our current knowledge about protein complexes is considerably better in

201 yeast than in human, with several real human complex components still missing the

202 respective functional annotations and thus leading to less homogeneous complex

203 predictions. Nevertheless, in most cases we tested, using NetAligner led to a higher

204 fraction of homogeneous complex predictions than a simple orthologs approach (Fig. 3).

205 In addition, we tested whether the annotations of the predicted complexes hint towards

206 specific functional roles in the cell by computing the statistical significance of GO term

207 enrichments in each complex (see *Materials & Methods*). We found that 233 of our

208 complex predictions in yeast (91%) and 87 in human (98%) were significantly

209 functionally enriched with respect to the given species interactome, indeed suggesting

210 specific cellular roles for those complexes and further validating our predictions. For the

211 complete list of predicted complexes, see Supplemental Table 1, in which we annotated

212 the complexes with the GO terms that are shared by most complex components to

213 provide information about their possible biological roles and highlighted those functional

214 annotations that are significantly enriched in the given complex. We also created

215 subsets of high-confidence (HC) predictions (Supplemental Table 1), requiring each

216 member complex to be functionally homogeneous and originate from a homogeneous

217 query complex with which it shares at least one homogeneous GO term (see *Materials*

218 *& Methods*). This resulted in HC sets of 105 yeast and 61 human complex predictions

219 (Table 2 and Fig. 2A & 2C), with 98% and 97% of them, respectively, being significantly

220 functionally enriched with respect to the given species interactome. We found 107 non-

221 redundant novel components across 63 HC yeast complexes and 91 non-redundant

222 novel components in 41 HC human complexes (Table 2 and Fig. 2B & 2D). For instance,

223 we predicted the proteins MPPA, MPPB, QCR1 and QCR2 to form a complex in human

224 (Fig. 4A). All four proteins are orthologous to the alpha and beta subunits of the

225 Mitochondrial processing peptidase (MPP) complex in yeast, which is involved in the

226 maturation of mitochondrial proteins by proteolytic cleavage of the N-terminal

227 localization sequence [40]. NetAligner found the interaction between MPPA and MPPB

228 of the yeast query complex to be conserved between human MPPA and MPPB, as well

229 as between the QCR1 and QCR2 orthologs (also known as UCR-1 and UCR-2). In

230 addition, this interaction was predicted to be likely conserved between the other

231 components of the complex (Fig. 4A), but no subcomplex of any of the four proteins was

232 found in current databases. According to our *in silico* validation experiments, all

233 components of the predicted complex are involved in proteolysis and have

234 metalloendopeptidase activity, but two of them (MPPA and MPPB) localise to the

235 mitochondrial matrix, while the other two (QCR1 and QCR2) localise to the

236 mitochondrial inner membrane as core components of the Cytochrome bc1 complex.

237    Although it might be that the MPP and QCR subunits form two separate complexes *in*
238    *vivo*, a combined MPP/QCR complex might also exist, since the two subcellular
239    localizations are adjacent, and it was observed that in plants, the MPP complex is
240    actually integrated into the cytochrome bc1 complex, with QCR1 and QCR2 being
241    identical to MPPB and MPPA, respectively [40]. Another interesting example is the
242    alignment of the human EXO1-MLH1-PCNA complex, which is involved in DNA-
243    mismatch repair, to the yeast interactome (Fig. 4B). The yeast complex predicted by the
244    alignment solution consists of six different proteins, PCNA, RAD27, DIN7, EXO1, MLH1
245    and MLH3, based on interactions existing either in human or yeast and predicted to be
246    likely conserved in the other species (Fig. 4B). The DNA sliding clamp PCNA and the
247    endonuclease RAD27 (also known as FEN1) are known to form a complex in DNA
248    replication and repair [41], and the interaction between MLH1 and MLH3 plays an
249    important role in meiotic recombination and mismatch repair [42]. In addition, the
250    double-stranded DNA exonucleases EXO1 and DIN7, both participating in mismatch
251    repair, have high sequence similarity, and the double knockout of EXO1 and RAD27 is
252    lethal [43]. Together, these findings point towards the possibility of a six-component
253    mismatch-repair complex in yeast and also that the corresponding human complex
254    might have additional components.

255    Finally, comparing the distributions of the total number of predicted (Fig. 2A & 2C) and
256    novel components (Fig. 2B & 2D), we found that both achieve higher values in yeast
257    than in human. However, this likely originates from the set of human query complexes
258    simply being considerably larger (1027 complexes) than the set of yeast query
259    complexes (244 complexes) and thus leading to both more complex predictions and a
260    higher total number of novel components in yeast. On the other hand, the exponential
261    decrease in the number of novel yeast complex components we observed (Fig. 2B)
262    compared to the considerably broader distribution of novel human complex components
263    (Fig. 2D), supports the view that our knowledge about yeast complexes surpasses that
264    of human ones. This is because, relative to complex predictions in yeast, a larger
265    fraction of predicted human complexes contains a given number of novel components.
266    For instance, only 24 out of 257 predicted yeast complexes (9.3%) contain three novel
267    components, while the same is true for 16 out of 89 predicted human complexes (18%).

### Predicting fly complexes from human and yeast data is complementary rather than redundant

The fruitfly *Drosophila melanogaster* is an important model organism. However, there does not yet exist a dedicated database of protein complexes in fly, the first large-scale screen, reporting 556 complexes, has only recently been completed [9], and only 221 complexes are annotated in GO [39]. We thus predicted protein complexes in that species by aligning the non-redundant sets of 244 yeast complexes and 1027 human complexes to the fly interactome (see *Materials & Methods*). This resulted in 66 non-redundant significant complex predictions originating from yeast and 219 from human (Table 3 and Fig. 2E & 2G), with only little overlap (see below). Compared to the set of 777 known complexes in fly (based on GO annotations and the recent large-scale screen by Guruharsha *et al.* [9]; see *Materials & Methods*), our complex predictions contained 252 non-redundant novel components across 56 complexes based on yeast data and 454 non-redundant novel components in 200 complexes originating from human data (Table 3 and Fig. 2F & 2H). Among those complexes, 6 and 45, respectively, were entirely novel. Again based on the recall and precision of the method (Table 1), we can estimate that most of the predicted complexes are probably sub-complexes and that at least 54% of the novel components we found are real complex members. The independent *in silico* validation (see *Materials & Methods*) revealed that the majority of the predicted fly complexes are functionally homogeneous, independent of whether they originated from yeast (70%) or human (68%) query complexes (Fig. 3C & 3D). This indicates that both of those organisms represent interesting sources for predicting evolutionary conserved protein complexes in fly based on network alignment. The fraction of homogeneous predicted fly complexes also lies between that of the human (63%) and yeast (97%) query complexes, indicating that our predictions are, on average, of at least the same quality as the manually-curated human complexes stored in the CORUM database [2], but of lower quality than the manually-curated yeast complexes in MPACT [1]. Here, we observed the biggest loss of homogeneity in the cellular component category (Fig. 3C). This might, however, result from the respective fly proteins missing sub-cellular annotations, which would again suggest that the current knowledge about protein complexes in yeast surpasses that of complexes in other species. Using a simple orthologs approach instead of NetAligner again leads to lower fractions of homogeneous complex predictions, independent of whether they originated

from yeast (58%) or human (60%) query complexes. Among the fly complex predictions, 58 based on yeast (88%) and 189 based on human data (86%) were significantly functionally enriched with respect to the fly interactome (see *Materials & Methods*), indicating specific cellular roles for those complexes and further validating our predictions. For the complete list of all predicted fly complexes, see Supplemental Table 2. The HC subsets of fly complex predictions (see *Materials & Methods* and Supplemental Table 2) consist of 42 fly complexes originating from yeast and 71 from human data (Table 3 and Fig. 2E & 2G), with 95% and 94% of them, respectively, being significantly functionally enriched with respect to the fly interactome. We found 134 non-redundant novel components across 32 HC complexes predicted from yeast and 168 non-redundant novel components in 62 HC complexes originating from human query complexes, with 4 and 11 of those complexes being completely novel (Table 3 and Fig. 2F & 2H). For instance, aligning the yeast Replication factor C (RFC) complex, consisting of the components RFC1-5, to the fly interactome predicted the corresponding fly assembly to encompass six proteins, Q9VKW3, Q9VX15, RFC1, RFC2, Q7KLW6 and Q9U9Q1 (Supplemental Fig. 1). According to the complex annotations in GO [39], the first four of those are known to belong to the RFC complex in fly, while the latter two are novel. Our alignments showed that both Q7KLW6 and Q9U9Q1 are orthologous to RFC2-5 in yeast, and our *in silico* validation experiments revealed that all six predicted components are involved in DNA replication and have ATP-binding capability, important for performing the loading of the DNA sliding clamp [44,45]. Moreover, in the HTP screen of Guruharsha *et al.* [9], the first five of those proteins were purified together, which provides further evidence for Q7KLW6 to actually be a component of the RFC complex in fly. On the other hand, this is also an example where NetAligner (through the incorporation of interaction data) was able to filter out false positive components that a simple orthologs approach would have predicted to belong to the complex. These comprise the proteins Q8T3K3, Q8IQ05 and Q95WV5, which are annotated as DNA replication accessory factors and most similar to the yeast chromosome transmission fidelity protein 18 (CTF18), which is known to substitute RFC1 in the RFC-like complex that also contains the proteins CTF8 and DCC1 and is required for establishment of chromosome cohesion in the S-phase of the cell cycle [46,47]. Moreover, the SNX subcomplex of the human Retromer complex, which is involved in mediating endosome to trans-Golgi network retrograde transport [48,49],

334 represents an example for predicting fly complexes based on human data (Fig. 4C). It

335 consists of the SNX1-SNX6 dimer that is important for membrane-bound coat formation

336 [48,49], and we predict this complex to exist in fly as an assembly of Q9VQQ6 and

337 Q9VLQ9. This prediction is not only supported by SNX1 and SNX6 being the best

338 BLAST [50] hits of Q9VQQ6 and Q9VLQ9, respectively, but also through the interaction

339 between the two human components being conserved in fly (Fig. 4C). Together, the

340 examples shown in Fig. 4 illustrate the fact that query complexes need not be fully

341 connected (i.e. can contain isolated proteins), but rather that NetAligner is indeed

342 capable of identifying conserved protein complexes despite the incompleteness of

343 current interactome networks [38].

344 Comparing the fly complex predictions from yeast and human data (Fig. 5), we found

345 only 26 pairs of matching complexes (i.e. pairs of predicted fly complexes with a

346 component overlap of more than 50% of each complex), covering 21 non-redundant

347 predicted complexes from human (10%) and 12 from yeast data (18%). One reason for

348 this little overlap between the fly complex predictions is probably the low number of

349 matching human and yeast query complexes. Indeed, only 41 (4%) and 46 (19%) of the

350 non-redundant sets of 1027 human and 244 yeast query complexes, respectively, are

351 present in the non-redundant set of matching human/yeast complex pairs. Between the

352 HC subsets, there are only 11 pairs of matching complexes, covering 11 non-redundant

353 predicted HC complexes from human (15%) and 7 from yeast data (17%; Fig. 5).

354 Moreover, none of the completely novel complexes we predicted was found based on

355 both yeast and human data. This clearly indicates that predicting fly complexes from

356 yeast and human query complexes through network alignment is complementary rather

357 than redundant. Protein complexes that were found both when starting from human and

358 from yeast data include the Replication factor C complex, the Casein kinase II, the 20S

359 core and 19/22S regulatory particles of the proteasome, as well as the Septin, Tubulin

360 and Actin filament complexes (Fig. 5), which all represent well-studied conserved

361 assemblies.

### *Mycoplasma complexes differ substantially from those of eukaryotes*

363 Kühner *et al.* [8] reported the first genome-wide analysis of protein complexes in the

364 human pathogen *Mycoplasma pneumoniae*, which has one of the smallest known

365 genomes (689 protein-encoding genes). This analysis revealed a rather complicated

366   machinery of almost 200 protein complexes, of which the majority were yet unknown [8].

367   To predict whether some of these are actually conserved in other organisms, we aligned

368   the non-redundant set of 174 mycoplasma complexes to the interactomes of yeast, fly

369   and human (see *Materials & Methods*). The complex to interactome alignments led to

370   only 11, 9 and 6 non-redundant significant predictions in those species, respectively

371   (Table 4). Compared to the sets of known protein complexes (see *Materials & Methods*),

372   our predictions contained 86 non-redundant novel components across 6 yeast

373   complexes, 68 across 9 fly complexes and 30 in 5 human complexes (Table 4). Based

374   on the recall and precision of the method (Table 1), we can again estimate that most of

375   the predicted complexes are probably sub-complexes and that at least 54% of the novel

376   components we found are real complex members. According to our independent *in silico*

377   validation (see *Materials & Methods*), only 55%, 44% and 33% of all predicted yeast, fly

378   and human complexes, respectively, are functionally homogeneous. If the mycoplasma

379   interactome (on which the query complexes are based) contained many false positive

380   interactions, one potential reason for the observed low functional homogeneity could be

381   that these complex predictions are less reliable. Other possible reasons include inherent

382   differences in the complexes repertoires of those species or aspects of biology that are

383   less well studied. In contrast, the fact that only 39% of all query mycoplasma complexes

384   are homogeneous is likely due to a lack of functional annotations in that organism.

385   Overall, since we used only the high-confidence subset of mycoplasma interactions (to

386   reduce the number of false positives), we think the very low numbers of significant

387   complex predictions (independent of the target species) and their low functional

388   homogeneity suggest that the mycoplasma interactome (at least the part currently

389   known) is very different from the interactomes of yeast, fly and human. This indicates

390   that there probably exist protein complexes that are unique to the pathogen and might

391   thus be targeted by drugs without causing adverse effects in the human host.

392   All complex predictions, except for one in yeast, were significantly functionally enriched

393   with respect to the given species interactome, suggesting that they perform specific

394   biological roles. For the complete list of predicted yeast, fly and human complexes,

395   based on mycoplasma data, see Supplemental Table 3. The HC subsets of those

396   complex predictions (see *Materials & Methods* and Supplemental Table 3), consist of

397   only three predicted complexes in yeast (one novel component), one in fly (four novel

398   components) and one in human (no novel components; Table 4), with all of them being

significantly functionally enriched with respect to the given species interactome. The mycoplasma query complexes that were predicted to be conserved in those HC sets comprise the DNA polymerase III complex in human, the ATP synthase complex in fly and yeast, as well as the 6-phosphofructokinase and the ribonucleoside-diphosphate reductase in yeast. The DNA polymerase III complex, consisting of DPO3X and the yet uncharacterized proteins Y007 and Y450, represents an interesting case: it was aligned to the RFC complex in human, whose clamp loading function is actually incorporated into the DNA polymerase III holoenzyme complex in prokaryotic species such as mycoplasma [51].

Comparing the different subsets of mycoplasma complexes that led to significant predictions in the other species (Fig. 6), one complex, the Restriction enzyme complex, could be aligned to protein complexes in all three species interactomes, two additional complexes were found in both yeast and human (the DNA polymerase III and Cohesin-like complexes), two in both fly and human (the Protein chaperone and Pyruvate dehydrogenase complexes), and four complexes were found in both the yeast and fly interactome (the ATP synthase, Peptidase, Ribosome and RNA polymerase complexes; Fig. 6). This means that the majority of mycoplasma complexes that led to significant predictions (56%) could be found in more than one target species, representing complexes involved in core biological processes conserved from bacteria up to higher eukaryotes. The alignment of the Restriction enzyme complex was, at first, surprising, because it is known to exist only in bacteria and archaea. A closer inspection, however, revealed that it was aligned to ATP-dependent metalloproteases in the eukaryotic species, suggesting that targeting the Restriction enzyme complex with drugs might also affect ATP-dependent metalloproteases in the human host and thus lead to potential adverse effects.

## Conclusions

Protein complexes represent key molecular entities that are implicated in many important biological processes within a cell. However, complexes are yet largely uncharacterized in most species and experimental determination of their composition is still a costly endeavour. Increasing our knowledge about protein complexes in important

model organisms via complex prediction thus represents an attractive option. Here, we showed that incorporating PPI data through network alignment significantly increases the precision of orthology-based complex prediction, though at the expense of missing some real complex components. By aligning known protein complexes to the interactomes of human, yeast and fly (the species with the highest number of interactions currently available in public databases) using NetAligner [36], our recently developed tool for network alignment, we were able to identify conserved protein (sub)complexes between human and yeast, as well as novel complex components, with higher precision than by using a simple orthologs-based approach. In addition, we predicted novel macromolecular assemblies (not present in current databases) in fly based on known yeast and human complexes, and found that our contemporary knowledge of yeast complexes surpasses that of other species, which is not surprising given the huge efforts invested into detecting macromolecular assemblies in yeast [3,4,5,6,7]. On the other hand, we found that current human and yeast complex data leads to complementary predictions in fly, meaning that querying known macromolecular assemblies of those species in the fly interactome unveils different conserved complexes. In the *in silico* validation experiments that we performed, our NetAligner-based complex predictions had about the same functional homogeneity as the known complexes we used for querying, and we were able to assign specific functional roles to almost all complexes. In contrast, predictions based on simple orthology mapping often resulted in reduced functional homogeneity. Finally, aligning the protein complexes of the human pathogen *M. pneumoniae* [8] to the interactomes of human, yeast and fly yielded only a handful of significant complex predictions, suggesting that the mycoplasma interactome is at least in parts quite different from those of eukaryotic species. Quantifying those differences is very difficult due to the incompleteness of current interactome data [38]. Nevertheless, since this issue is at least partly addressed through the prediction of likely conserved interactions by the NetAligner algorithm, it indicates that some of the protein complexes that have been identified in mycoplasma might not have any counterpart in eukaryotes and thus represent interesting drug targets with little chances of causing adverse effects in the human host. On the other hand, our analyses revealed that the pathogen's Restriction enzyme complex, which is known to be unique to bacteria and archaea, is similar to ATP-dependent metalloproteases in human, implying that targeting this complex might have undesirable outcomes.

## Materials & Methods

### *Datasets of protein sequences*

We collected datasets of protein sequences for human (*Homo sapiens*), fly (*Drosophila melanogaster*), yeast (*Saccharomyces cerevisiae*) and mycoplasma (*Mycoplasma pneumoniae*) from the UniProt Knowledgebase release 15.8 [52] by merging the set of sequences stored in Swiss-Prot (including splice variants) and TrEMBL with experimental evidence on protein or transcript level. For mycoplasma, due to a lack of annotation data, we also included sequences not yet marked as having experimental evidence, but excluded all sequences that were only present in TrEMBL and thus of low confidence. Clustering based on 100% sequence identity resulted in non-redundant sets of 75,981 human, 23,296 fly, 6,121 yeast and 687 mycoplasma protein sequences.

### *Lists of orthologous proteins*

We determined lists of orthologous proteins for species combinations of human, fly, yeast and mycoplasma by performing a reciprocal BLASTP [50] search. We used an E-value threshold of $10^{-10}$ and considered only hits in the top10 of the BLASTP output to filter out spurious matches. This resulted in non-redundant sets of 91,112 human/fly, 19,558 human/yeast, 12,778 fly/yeast, 1,005 human/mycoplasma, 644 fly/mycoplasma and 488 yeast/mycoplasma orthologs.

### *Interactome construction*

We built interactome networks for human, fly and yeast by extracting and merging binary protein-protein interaction data from the major interaction databases IntAct [53], MINT [54] and HPRD (for human) [55]. To increase the quality of the resulting binary interactome networks, we then filtered out all those interactions without support in form of Pubmed ID(s) or interaction detection method(s). For mycoplasma, we extracted the list of high-confidence, binary protein-protein interactions from Kühner *et al.* [8], mapping ordered locus names to UniProt accession codes [52]. This resulted in non-redundant

488  interactomes consisting of 53,290 interactions in human, 19,260 in fly, 60,721 in yeast

489  and 1,058 in mycoplasma (Supplemental Table 4).

490  ### Non-redundant benchmark set of human/yeast complex pairs

491  We used the non-redundant benchmark set of conserved human/yeast complex pairs

492  described in [36]. In brief, this set is based on the manually-curated yeast complexes

493  from the MPACT database [1] and human complexes from the CORUM database [2]

494  whose components are fully present in the respective species interactome, and

495  clustered to remove redundancy. Matching (i.e. conserved) complexes between the two

496  species are defined by requiring at least 2 and 25% of the components of the given

497  human complex to have at least one ortholog in the respective yeast complex and vice

498  versa. The complete benchmark set encompasses 71 matching human/yeast complex

499  pairs, consisting of 64 non-redundant human and 52 non-redundant yeast complexes.

500  ### Complex to interactome alignment using NetAligner

501  We performed all complex to interactome alignments using the NetAligner algorithm that

502  we recently developed [36] and which is also available as a web server [56]. For all

503  species combinations, we computed vertex and interaction conservation probabilities

504  required by NetAligner based on the interactomes and lists of orthologous proteins as

505  described in [36], using default parameters. We assigned reliabilities to each interaction

506  based on the number of Pubmed IDs supporting it as previously described [26]. For

507  aligning the complexes to whole organism interactomes, we created a network

508  representation of each complex, consisting of all interactions between complex

509  components that are present in the respective species interactome and including self-

510  interactions with a reliability of 0 for all singletons in order to not lose any information

511  about complex composition [36].

512  ### Performance comparison to simple orthologs approach

513  To compare the performance of NetAligner to that of a simple orthologs approach, in

514  which the set of all orthologs of the components of a given query complex are predicted

515  to be part of the complex in the target species, we evaluated the results of this approach

516  with respect to the benchmark set of protein complexes as described in [36]. In brief, the

517  complex predictions are evaluated in terms of how well they agree with the respective

matching benchmark set complexes based on the overlap of their protein components. For each complex predicted by the orthologs approach, we first determined the best-matching benchmark complex of the same species by minimising the total number of unmatched components. A complex prediction was deemed to 'cover' a benchmark complex if it contained at least 2 and at least 50% of its components. We then calculated the number of true positives (TP) as the total number of distinct proteins common to any given complex prediction and the benchmark complex it covers; the number of false positives (FP) as the total number of distinct proteins that are part of any given complex prediction, but not present in the benchmark complex it covers; and the number of false negatives (FN) as the total number of distinct proteins present in any given benchmark complex, but not part of any complex prediction covering that complex. Next, we computed the performance of the orthologs approach in terms of precision and recall:

$$\text{precision} \;=\; \frac{\text{TP}}{\text{TP}+\text{FP}}; \quad \text{recall} \;=\; \frac{\text{TP}}{\text{TP}+\text{FN}}$$

Finally, we report the average precision and recall of predicting yeast complexes based on human protein complex data and vice versa to avoid parameter overfitting (Fig. 1B). For NetAligner, since we used the same performance evaluation strategy, as well as the same list of orthologs and set of benchmark complexes, we could directly take the performance results reported in our previous work for complex to interactome alignment [36]. We also evaluated the performance when using default parameters (Fig. 1B).

### Non-redundant sets of protein complexes in human, yeast and mycoplasma

We collected non-redundant sets of protein complexes in human, yeast and mycoplasma. For this, we first extracted all human complexes from the CORUM database [2], the set of manually curated yeast complexes from the MPACT database [1], as well as all homo- and heteromeric mycoplasma complexes from Kühner et al. [8]. Analogously to the procedure for constructing the non-redundant benchmark set of human/yeast complex pairs [36], we then filtered out those complexes that were not fully present in the respective species interactome and clustered them based on the overlap of their components using complete linkage hierarchical clustering to remove redundancy. The distance $d(c_1, c_2)$ between two complexes $c_1$ and $c_2$ was defined as:

$$d(c_1, c_2) = 1 - \frac{|c_1 \cap c_2|}{\max(|c_1|, |c_2|)},$$

and we cut the resulting dendrogram using a distance threshold of 0.5, such that each pair of complexes in the same cluster share more than 50% of their components (choosing the largest complex of each cluster as its representative). This resulted in non-redundant sets of 1027 protein complexes in human, 244 in yeast and 174 in mycoplasma.

### Identification of novel complex components

To identify novel components in our complex predictions, we compared them with the set of known complexes of the respective species. We took all 1826 known human complexes from the CORUM database [2], all 402 human complexes annotated in the Gene Ontology (GO) [39], as well as the 155 and 622 complexes from the recent high-throughput (HTP) studies of Hutchins *et al.* [10] and Havugimana *et al.* [11], respectively (3005 complexes in total). For yeast, we collected all 263 manually curated and all 871 HTP complexes from the MPACT database [1], which include the large-scale studies performed by Gavin *et al.* [3] and Ho *et al.* [4]. We then added the 491 and 547 complexes defined in the HTP studies by Gavin *et al.* [5] and Krogan *et al.* [6], respectively, as well as all 350 yeast complexes annotated in GO and the recently published set of 720 yeast complexes from Babu *et al.* [7] (3242 complexes in total). Since there does not yet exist a dedicated database of protein complexes in fly, we determined the set of known fly complexes based on the set of HTP complexes reported by Guruharsha *et al.* [9] and shared GO annotations (child terms of 'macromolecular complex' (GO:0032991)), similar to [57], resulting in a total of 777 fly protein complexes. For each predicted complex in a given organism, we then determined the known complex of that species with the largest overlap in terms of protein components and marked all those proteins as novel that are not part of the known complex.

### Computation of functional homogeneity

We computed the functional homogeneity of protein complexes as an automated strategy to validate their composition in terms of protein components. First, we extracted the GO protein annotations from the UniProt database [52] for all GO categories (i.e.

576 biological process, molecular function and cellular component). Then, we calculated the

577 GO homogeneity $h(c)$ of each complex $c$ per GO category, defined as the maximum

578 fraction of protein components $p(c)$ that share the same GO term $t$ [58]:

$$h(c) \quad = \quad \max_t \frac{|p_t(c)|}{|p(c)|}$$

579 We classified each complex with a GO homogeneity of higher than 0.5 as functionally

580 homogeneous in the given GO category. Lastly, to increase the confidence level of all

581 subsequent analyses, we defined all those complexes as functionally homogeneous that

582 fulfilled this criterion in at least two GO categories.

583 ### *Statistical significance of functional enrichments*

584 For all most-abundant functional annotations of a given complex and GO category (i.e.

585 those that contribute to its functional homogeneity in that category), we determined the

586 statistical significance of their enrichments in the complex with respect to the given

587 species interactome, using a one-sided Fisher's exact test with Bonferroni multiple

588 testing correction and a strict p-value threshold of 0.025. We then defined all those

589 complexes as significantly functionally enriched that had a significant enrichment p-

590 value in at least two GO categories.

591 ### *High-confidence subsets of complex predictions*

592 We post-processed all sets of predicted complexes to define high-confidence (HC)

593 subsets. A predicted complex has to fulfil the following criteria to be present in the given

594 HC set: it has to (i) be part of the non-redundant subset of significant predictions, (ii)

595 originate from a functionally homogeneous query complex, (iii) be homogeneous itself

596 and (iv) share at least one homogeneous GO term with the given query complex. Query

597 complexes were defined as high-confidence if they were functionally homogeneous.

598 ## Acknowledgments

599 We would like to thank Sebastian Kühner (Anne-Claude Gavin's lab at EMBL

600 Heidelberg) for providing the mycoplasma complexes and interaction data, as well as

Amelie Stein (UCSF) and Andreas Zanzoni (University of Aix-Marseille) for helpful discussions.

# References

1. Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, Stümpflen V (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436-441.
2. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W (2010) CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 38: D497-501.
3. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier M-A, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147.
4. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CWV, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180-183.
5. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A-M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-636.
6. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637-643.
7. Babu M, Vlasblom J, Pu S, Guo X, Graham C, Bean BDM, Burston HE, Vizeacoumar FJ, Snider J, Phanse S, Fong V, Tam YYC, Davey M, Hnatshak O, Bajaj N, Chandran S, Punna T, Christopolous C, Wong V, Yu A, Zhong G, Li J, Stagljar I, Conibear E, Wodak

643         SJ, Emili A, Greenblatt JF (2012) Interaction landscape of membrane-protein complexes
644         in Saccharomyces cerevisiae. Nature 489: 585-589.
645 8. Kühner S, Van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T,
646         Bader S, Beltran-Alvarez P, Castano-Diez D, Chen W-H, Devos D, Guell M, Norambuena
647         T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Böttcher B, Frangakis
648         AS, Russell RB, Serrano L, Bork P, Gavin A-C (2009) Proteome Organization in a
649         Genome-Reduced Bacterium. Science 326: 1235-1240.
650 9. Guruharsha KG, Rual J-F, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee
651         DY, Cenaj O, McKillip E, Shah S, Stapleton M, Wan KH, Yu C, Parsa B, Carlson JW,
652         Chen X, Kapadia B, VijayRaghavan K, Gygi SP, Celniker SE, Obar RA, Artavanis-
653         Tsakonas S (2011) A protein complex network of Drosophila melanogaster. Cell 147: 690-
654         703.
655 10. Hutchins JRA, Toyoda Y, Hegemann B, Poser I, Hériché J-K, Sykora MM, Augsburg M,
656         Hudecz O, Buschhorn BA, Bulkescher J, Conrad C, Comartin D, Schleiffer A, Sarov M,
657         Pozniakovsky A, Slabicki MM, Schloissnig S, Steinmacher I, Leuschner M, Ssykor A,
658         Lawo S, Pelletier L, Stark H, Nasmyth K, Ellenberg J, Durbin R, Buchholz F, Mechtler K,
659         Hyman AA, Peters J-M (2010) Systematic Analysis of Human Protein Complexes
660         Identifies Chromosome Segregation Proteins. Science (New York, NY).
661 11. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong
662         V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar V-U-N, Bezginov A,
663         Clark GW, Wu GC, Wodak SJ, Tillier ERM, Paccanaro A, Marcotte EM, Emili A (2012) A
664         census of human soluble protein complexes. Cell 150: 1068-1081.
665 12. Dezso Z, Oltvai ZN, Barabási A-L (2003) Bioinformatics analysis of experimentally
666         determined protein complexes in the yeast Saccharomyces cerevisiae. Genome Research
667         13: 2450-2454.
668 13. Hart GT, Lee I, Marcotte ER (2007) A high-accuracy consensus map of yeast protein
669         complexes reveals modular nature of gene essentiality. BMC bioinformatics 8: 236.
670 14. Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC,
671         Krogan NJ, Koller D (2009) A complex-based reconstruction of the Saccharomyces
672         cerevisiae interactome. Molecular & Cellular Proteomics 8: 1361-1381.
673 15. Pache RA, Babu MM, Aloy P (2009) Exploiting gene deletion fitness effects in yeast to
674         understand the modular architecture of protein complexes under different growth
675         conditions. BMC Systems Biology 3: 74.
676 16. Pache RA, Aloy P (2008) Incorporating high-throughput proteomics experiments into
677         structural biology pipelines: identification of the low-hanging fruits. PROTEOMICS 8:
678         1959-1964.
679 17. Brooks MA, Gewartowski K, Mitsiki E, Létoquart J, Pache RA, Billier Y, Bertero M, Corréa
680         M, Czarnocki-Cieciura M, Dadlez M, Henriot V, Lazar N, Delbos L, Lebert D,
681         Piwowarski J, Rochaix P, Böttcher B, Serrano L, Séraphin B, van Tilbeurgh H, Aloy P,
682         Perrakis A, Dziembowski A (2010) Systematic bioinformatics and experimental validation
683         of yeast complexes reduces the rate of attrition during structural investigations. Structure
684         (London, England : 1993) 18: 1075-1082.
685 18. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks.
686         Proceedings of the National Academy of Sciences of the United States of America 100:
687         12123-12128.
688 19. Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from
689         protein interaction networks. Proteins 54: 49-57.
690 20. Poyatos JF, Hurst LD (2004) How biologically relevant are interaction-based modules in
691         protein networks? Genome biology 5: R93.

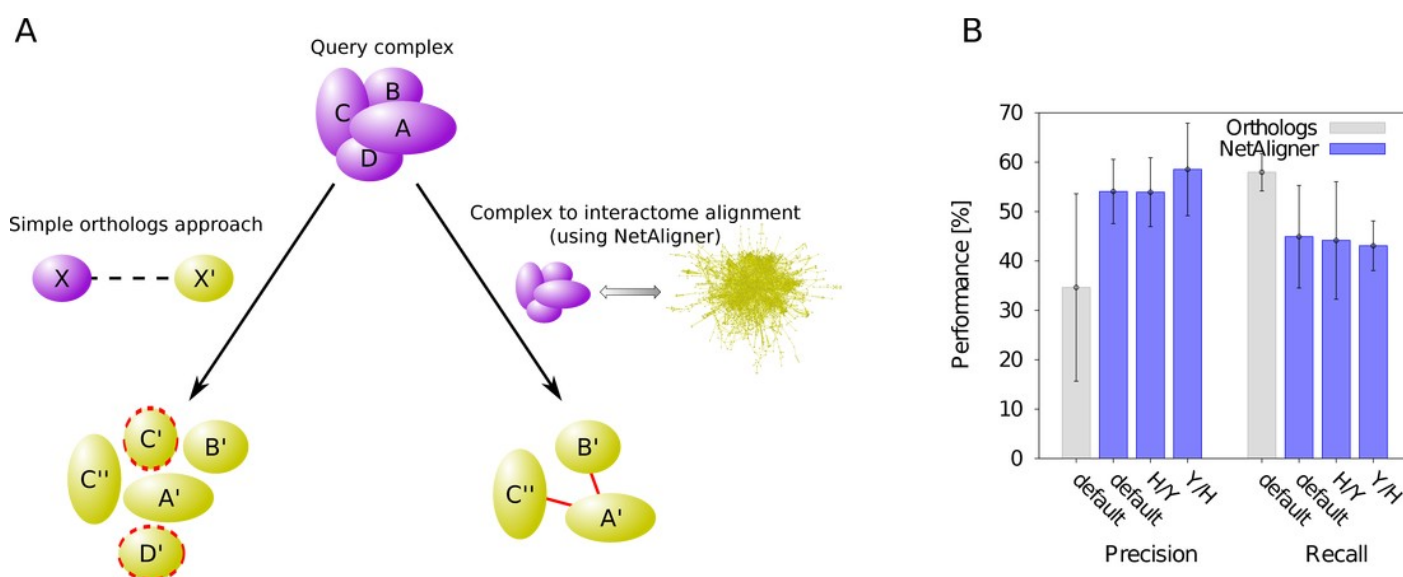
692 21. Bader GD, Hogue CWV (2003) An automated method for finding molecular complexes in
693      large protein interaction networks. BMC bioinformatics 4: 2.
694 22. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ (2007) Identifying functional modules in
695      the physical interactome of Saccharomyces cerevisiae. Proteomics 7: 944-960.
696 23. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M,
697      Park H-O, Hayles J, Hoe K-L, Kim D-U, Ideker T, Grewal SI, Weissman JS, Krogan NJ
698      (2008) Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map
699      in Fission Yeast. Science 322: 405-410.
700 24. van Dam TJP, Snel B (2008) Protein complex evolution does not involve extensive network
701      rewiring. PLoS Computational Biology 4: e1000132.
702 25. Koonin EV, Wolf YI, Aravind L (2001) Prediction of the archaeal exosome and its
703      connections with the proteasome and the translation and transcription machineries by a
704      comparative-genomic approach. Genome Research 11: 240-252.
705 26. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T (2003) Conserved
706      pathways within bacteria and yeast as revealed by global protein network alignment. Proc
707      Natl Acad Sci USA 100: 11394-11399.
708 27. Sharan R, Ideker T, Kelley B, Shamir R, Karp RM (2005) Identification of protein complexes
709      by comparative analysis of yeast and bacterial protein interaction data. Journal of
710      computational biology : a journal of computational molecular cell biology 12: 835-846.
711 28. Koyutürk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A (2006) Pairwise
712      alignment of protein interaction networks. J Comput Biol 13: 182-199.
713 29. Sharan R, Ideker T (2006) Modeling cellular machinery through biological network
714      comparison. Nat Biotechnol 24: 427-433.
715 30. Cootes AP, Muggleton SH, Sternberg MJE (2007) The identification of similarities between
716      biological networks: application to the metabolome and interactome. Journal of Molecular
717      Biology 369: 1126-1139.
718 31. Kiemer L, Cesareni G (2007) Comparative interactomics: comparing apples and pears?
719      Trends Biotechnol 25: 448-454.
720 32. Narayanan M, Karp RM (2007) Comparing protein interaction networks via a graph match-
721      and-split algorithm. J Comput Biol 14: 892-907.
722 33. Hirsh E, Sharan R (2007) Identification of conserved protein complexes based on a model of
723      protein network evolution. Bioinformatics 23: e170-176.
724 34. Ali W, Deane CM (2009) Functionally guided alignment of protein interaction networks for
725      module detection. Bioinformatics 25: 3166-3173.
726 35. Dost B, Shlomi T, Gupta N, Ruppin E, Bafna V, Sharan R (2008) QNet: a tool for querying
727      protein interaction networks. Journal of Computational Biology 15: 913-925.
728 36. Pache RA, Aloy P (2012) A novel framework for the comparative analysis of biological
729      networks. PloS one 7: e31220.
730 37. Mosca R, Pache RA, Aloy P (2012) The role of structural disorder in the rewiring of protein
731      interactions through evolution. Molecular & Cellular Proteomics: 1-8.
732 38. Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T,
733      Zenkner M, Xin X, Goh K-I, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie
734      JM, Cevik S, Simon C, de Smet A-S, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D,
735      Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone
736      C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási A-L, Vidal
737      M (2009) An empirical framework for binary interactome mapping. Nat Methods 6: 83-
738      90.
739 39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K,
740      Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese

741 JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for
742 the unification of biology. The Gene Ontology Consortium. Nature genetics 25: 25-29.
743 40. Nomura H, Athauda SBP, Wada H, Maruyama Y, Takahashi K, Inoue H (2006) Identification
744 and reverse genetic analysis of mitochondrial processing peptidase and the core protein of
745 the cytochrome bc1 complex of Caenorhabditis elegans, a model parasitic nematode.
746 Journal of biochemistry 139: 967-979.
747 41. Gomes XV, Burgers PM (2000) Two modes of FEN1 binding to PCNA regulated by DNA.
748 The EMBO journal 19: 3811-3821.
749 42. Wang TF, Kleckner N, Hunter N (1999) Functional specificity of MutL homologs in yeast:
750 evidence for three Mlh1-based heterocomplexes with distinct roles during meiosis in
751 recombination and mismatch correction. Proceedings of the National Academy of
752 Sciences of the United States of America 96: 13914-13919.
753 43. Tishkoff DX, Boerger AL, Bertrand P, Filosi N, Gaida GM, Kane MF, Kolodner RD (1997)
754 Identification and characterization of Saccharomyces cerevisiae EXO1, a gene encoding
755 an exonuclease that interacts with MSH2. Proceedings of the National Academy of
756 Sciences of the United States of America 94: 7487-7492.
757 44. Gomes XV, Schmidt SL, Burgers PM (2001) ATP utilization by yeast replication factor C. II.
758 Multiple stepwise ATP binding events are required to load proliferating cell nuclear
759 antigen onto primed DNA. The Journal of biological chemistry 276: 34776-34783.
760 45. Schmidt SL, Gomes XV, Burgers PM (2001) ATP utilization by yeast replication factor C. III.
761 The ATP-binding domains of Rfc2, Rfc3, and Rfc4 are essential for DNA recognition and
762 clamp loading. The Journal of biological chemistry 276: 34784-34791.
763 46. Mayer ML, Gygi SP, Aebersold R, Hieter P (2001) Identification of RFC(Ctf18p, Ctf8p,
764 Dcc1p): an alternative RFC complex required for sister chromatid cohesion in S.
765 cerevisiae. Molecular cell 7: 959-970.
766 47. Naiki T, Kondo T, Nakada D, Matsumoto K, Sugimoto K (2001) Chl12 (Ctf18) forms a novel
767 replication factor C-related complex and functions redundantly with Rad24 in the DNA
768 replication checkpoint pathway. Molecular and cellular biology 21: 5838-5845.
769 48. Wassmer T, Attar N, Bujny MV, Oakley J, Traer CJ, Cullen PJ (2007) A loss-of-function
770 screen reveals SNX5 and SNX6 as potential components of the mammalian retromer.
771 Journal of cell science 120: 45-54.
772 49. Hong Z, Yang Y, Zhang C, Niu Y, Li K, Zhao X, Liu J-J (2009) The retromer component
773 SNX6 interacts with dynactin p150(Glued) and mediates endosome-to-TGN transport.
774 Cell research 19: 1334-1349.
775 50. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997)
776 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
777 Nucleic Acids Res 25: 3389-3402.
778 51. Pomerantz RT, O'Donnell M (2007) Replisome mechanics: insights into a twin DNA
779 polymerase machine. Trends in microbiology 15: 156-164.
780 52. UniProt-Consortium (2009) The Universal Protein Resource (UniProt) 2009. Nucleic acids
781 research 37: D169-174.
782 53. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M,
783 Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M,
784 Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K,
785 Hermjakob H (2010) The IntAct molecular interaction database in 2010. Nucleic Acids
786 Research 38: D525-531.
787 54. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L,
788 Cesareni G (2010) MINT, the molecular interaction database: 2009 update. Nucleic Acids
789 Research 38: D532-539.

790 55. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S,
791   Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A,
792   Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S,
793   Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA,
794   Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human Protein
795   Reference Database--2009 update. Nucleic Acids Res 37: D767-772.
796 56. Pache RA, Céol A, Aloy P (2012) NetAligner--a network alignment server to compare
797   complexes, pathways and whole interactomes. Nucleic acids research.
798 57. Bruckner S, Hüffner F, Karp RM, Shamir R, Sharan R (2010) Topology-free querying of
799   protein interaction networks. J Comput Biol 17: 237-252.
800 58. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) The human disease
801   network. Proc Natl Acad Sci USA 104: 8685-8690.

# Figure 1

Network alignment vs. simple orthologs approach in predicting protein complexes.

Comparison of network alignment (using NetAligner [36] ) and a simple orthologs approach in predicting protein complexes in a target species based on known complexes in a query organism. A: In the simple orthologs approach, all proteins in the target species (yellow) that are orthologous to the proteins of the query complex (violet) are predicted to form the given complex in the target organism. This can lead to false positives (proteins marked with red dashed lines) that are not part of the real complex. In the more sophisticated network alignment strategy, only those orthologs are predicted to be part of the given complex in the target species that are supported by conserved or likely-conserved interactions (red) between the two organism interactomes. B: Performance of the orthologs (grey) and NetAligner (blue) methods on a benchmark set of known conserved human/yeast complex pairs, using default parameters (default) or parameter sets trained on one species pair (H/Y, Y/H) and evaluated across both species pairs. Error bars denote one standard error of the mean.

NetAligner parameters and expected alignment performance.

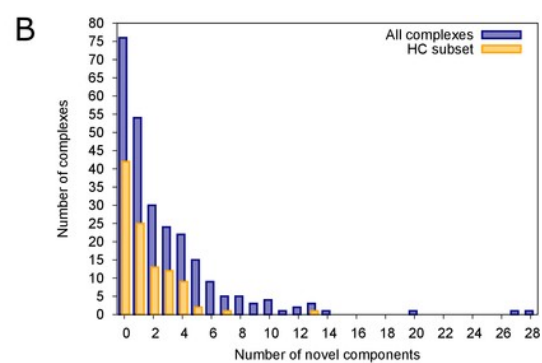
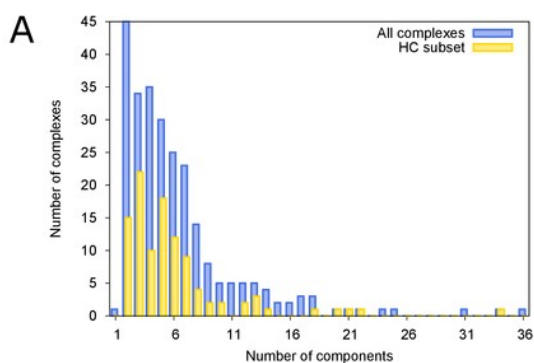
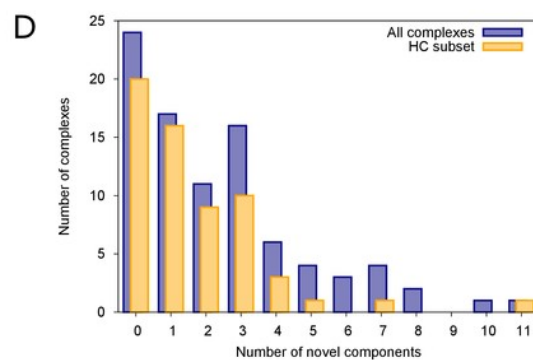
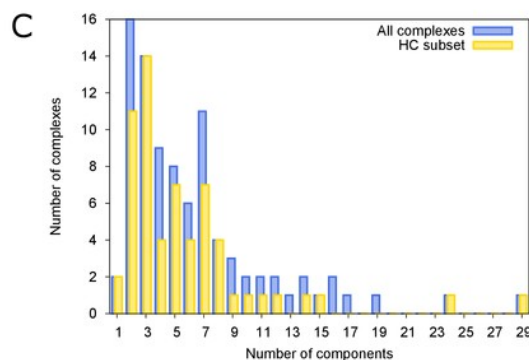
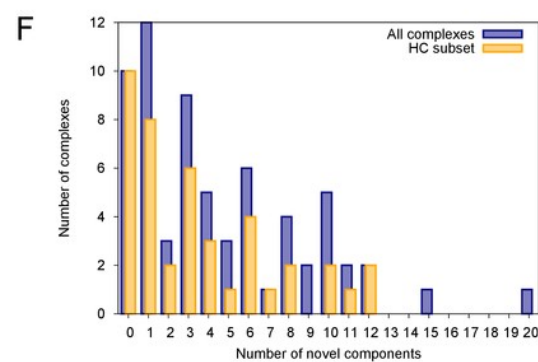
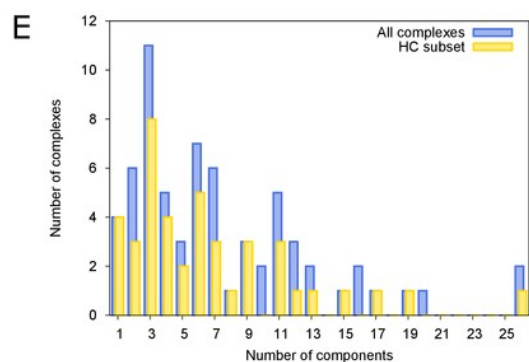
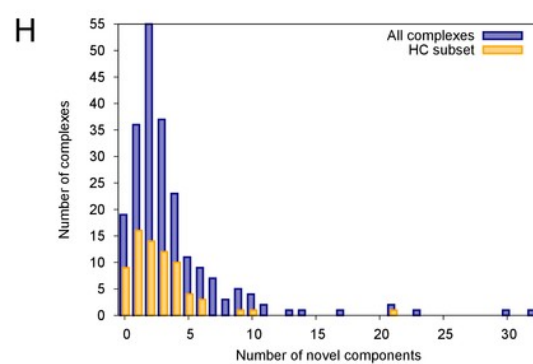
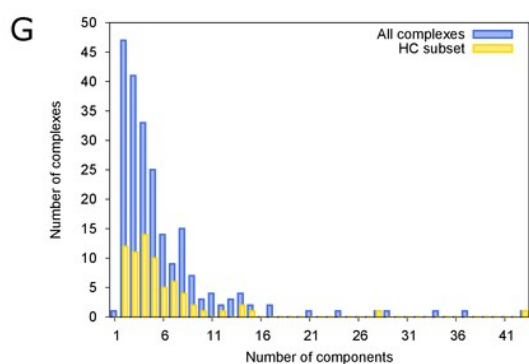
The NetAligner parameters for complex to interactome alignment were taken from [36]. For yeast to human and human to yeast alignments, we took the best performing parameter combinations for these species as determined in the benchmarks [36]. For all other species, we used the default parameters for complex to interactome alignment. The option to predict likely conserved interactions is always set to true (1), because this considerably improves alignment performance [36]. Precision and recall describe the expected performance of NetAligner in correctly identifying protein complex components (see Materials & Methods).

| | | Complex to interactome alignment | | |
|---|---|---|---|---|
| | | Yeast to human | Human to yeast | Other species |
| **Parameters** | Predict likely conserved interactions | 1 | 1 | 1 |
| | Vertex probability threshold | 0.0 | 0.0 | 0.0 |
| | Edge probability threshold | 0.1 | 0.0 | 0.0 |
| | Max insertion length | 1 | 2 | 2 |
| | Vertex to edge score balance | 0.2 | 0.0 | 0.1 |
| **Performance** | Precision [%] | 49.17 | 60.91 | 54.07 |
| | Recall [%] | 38.06 | 56.06 | 44.91 |

# Figure 2

Complex size and number of novel components distributions for all predicted complexes.

Distributions of the total number of components of all predicted complexes (blue) and the high-confidence (HC) subset (yellow) are shown on the left. Distributions of the number of novel components found in all predicted complexes (violet) and in the HC subset (orange) are shown on the right. A&B: prediction of yeast complexes based on human data; C&D: prediction of human complexes based on yeast data; E&F: prediction of fly complexes based on yeast data; G&H: prediction of fly complexes based on human data.

Human to yeast

Yeast to human

Yeast to fly

Human to fly

Complex predictions in yeast and human.

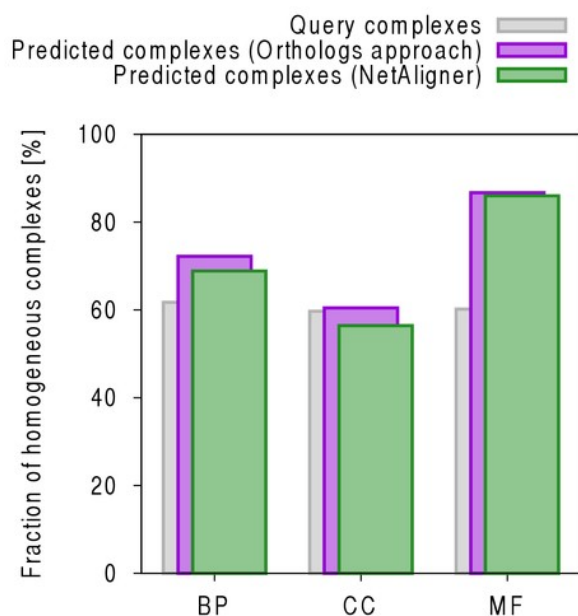
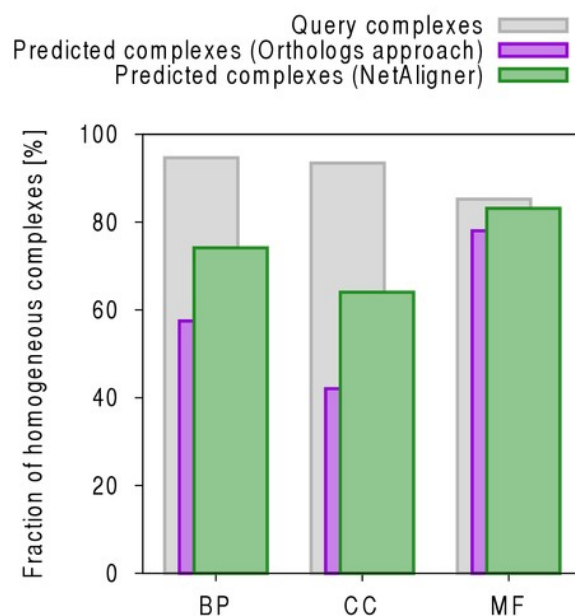
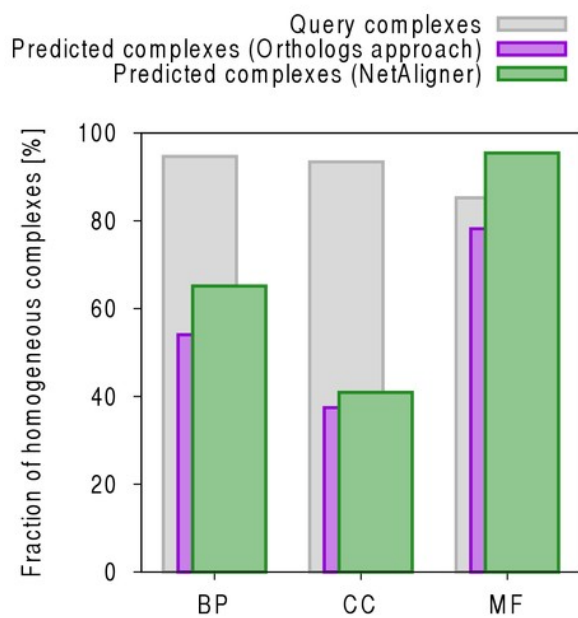
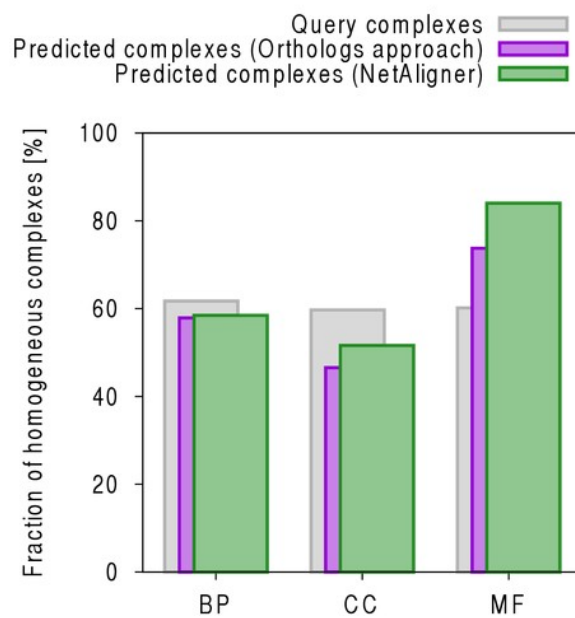
Basic statistics of the complex prediction results in yeast and human, based on aligning known human complexes to the yeast interactome and vice versa (see Materials & Methods). Results are shown both for all complex predictions (All) and for the high-confidence subset (HC). #, number of; nr, non-redundant.

| Prediction type | Human to yeast | | Yeast to human | |
|---|---|---|---|---|
| | All | HC | All | HC |
| # query complexes | 1027 | 645 | 244 | 236 |
| # predicted complexes (nr) | 257 | 105 | 89 | 61 |
| Total # complex components (nr) | 604 | 372 | 464 | 325 |
| Average # proteins per complex | 6.4 | 5.93 | 6.37 | 5.67 |
| Total # novel components (nr) | 307 | 107 | 175 | 91 |
| Average # novel components per complex | 2.75 | 1.46 | 2.4 | 1.62 |
| # entirely novel complexes (nr) | 0 | 0 | 2 | 2 |

# Figure 3

Functional homogeneity of query and predicted complexes.

Comparison of the functional homogeneity of query (grey) and predicted complexes, using either a simple orthologs approach (purple) or NetAligner (green), in the different Gene Ontology (GO) annotation categories. BP, Biological Process; CC, Cellular Component; MF, Molecular Function. A: prediction of yeast complexes based on human data; B: prediction of human complexes based on yeast data; C: prediction of fly complexes based on yeast data; D: prediction of fly complexes based on human data.

**A**  Human to yeast

**B**  Yeast to human

**C**  Yeast to fly

**D**  Human to fly

Prediction of protein complexes in fly.

Basic statistics of the complex prediction results in fly, based on aligning known yeast and human complexes to the fly interactome (see Materials & Methods). Results are shown both for all complex predictions (All) and for the high-confidence subset (HC). #, number of; nr, non-redundant.

| Prediction type | Yeast to fly | | Human to fly | |
|---|---|---|---|---|
| | All | HC | All | HC |
| # query complexes | 244 | 236 | 1027 | 645 |
| # predicted complexes (nr) | 66 | 42 | 219 | 71 |
| Total # complex components (nr) | 405 | 255 | 640 | 291 |
| Average # proteins per complex | 7.5 | 6.69 | 5.84 | 5.93 |
| Total # novel components (nr) | 252 | 134 | 454 | 168 |
| Average # novel components per complex | 4.61 | 3.55 | 3.69 | 2.79 |
| # entirely novel complexes (nr) | 6 | 4 | 45 | 11 |

# Figure 4

Examples for complex prediction based on network alignment.

Examples of using NetAligner [36] to predict protein complexes based on aligning network representations of known protein complexes in one species to the interactome of another species. Notably, NetAligner does not require query complexes to be fully connected. Edge colors in the alignment solutions show which species the given interaction that was predicted to be likely conserved originates from. Green edges denote known conserved interactions. High-confidence interactions are highlighted with thicker edges. Alignment solution nodes represent pairs of orthologous proteins (separated by '/') between the respective species, allowing both one-to-many and many-to-many orthology relationships. Components of the given predicted complex in the target species (extracted from the respective alignment solution) are listed, together with the GO biological process and molecular function annotations that could be assigned to the predicted complex. A: prediction of a mitochondrial processing peptidase complex in human; B: prediction of a mismatch repair complex in yeast; C: prediction of a SNX complex in fly. See main text for details. Network representations were created with Cytoscape [59].

Network representation of the query complex | NetAligner | Target interactome | Alignment solution | Predicted target complex

A

Yeast mitochondrial processing peptidase complex

Human

**Components:**
MPPA, MPPB, QCR1, QCR2

**Biological process:**
Proteolysis

**Molecular function:**
Metalloendopeptidase activity,
Zinc ion binding

B

Human EXO1-MLH1-PCNA mismatch repair complex

Yeast

**Components:**
EXO1, MLH1, MLH3, DIN7,
RAD27, PCNA

**Biological process:**
Mismatch repair

**Molecular function:**
no common annotation

C

Human SNX complex

Fly

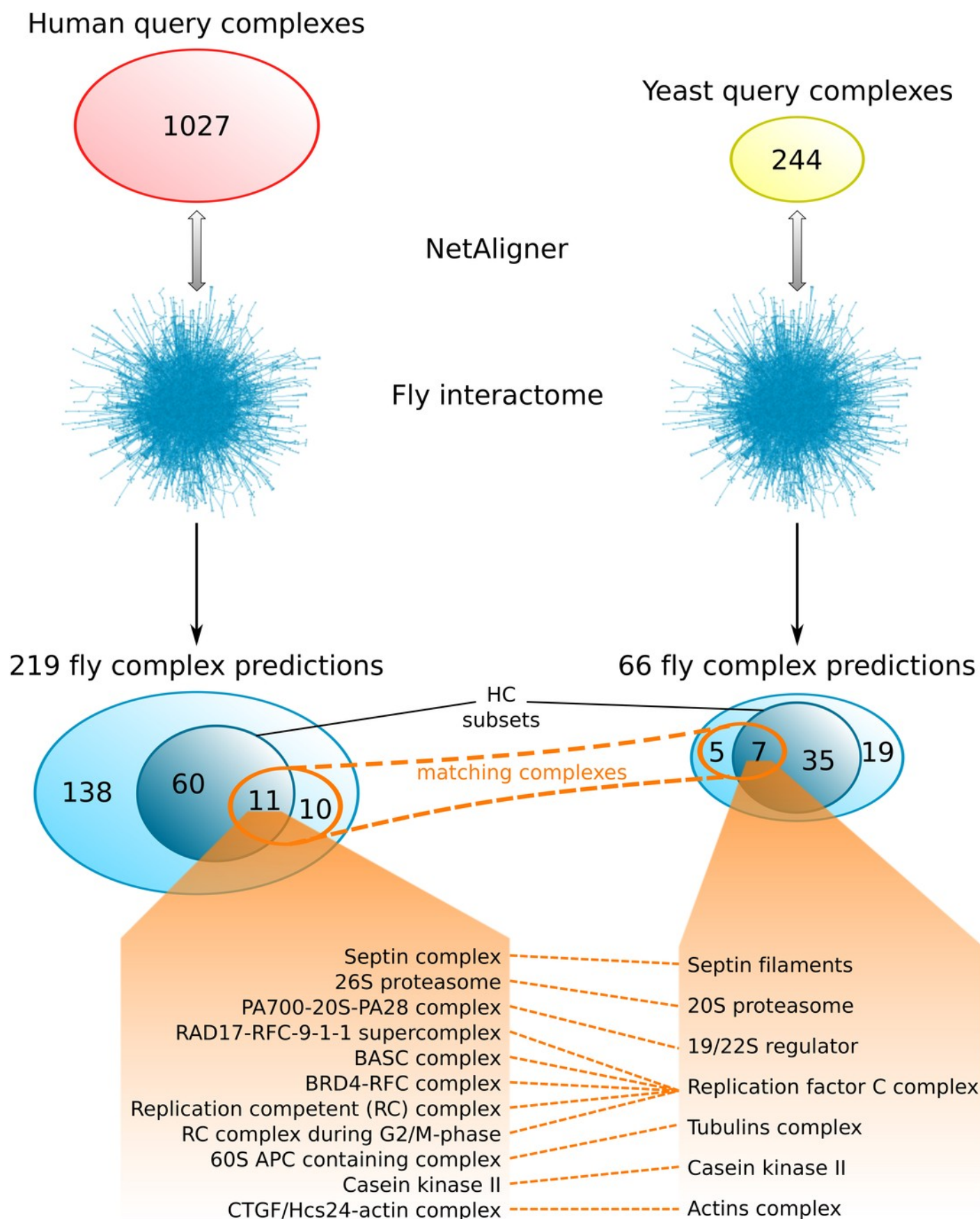
**Components:**
Q9VLQ9, Q9VQQ6

**Biological process:**
Cell communication

**Molecular function:**
Phosphoinositide binding

# Figure 5

Comparison of fly complex predictions originating from yeast and human data.

Comparison of the 219 and 66 non-redundant fly complexes predicted through alignment of human (red) and yeast (yellow) query complexes, respectively, to the fly interactome using NetAligner [36]. High-confidence (HC) subsets of the complex predictions are shown in darker blue. Matching complexes, i.e. those that were identified both from human and yeast query complexes with more than 50% shared components, are highlighted in orange. The 11 pairs of matching complexes between the two HC subsets are listed below.

Predictions based on known mycoplasma complexes.

Basic statistics of the complex prediction results in yeast, fly and human, based on aligning known mycoplasma complexes to the respective species interactome (see Materials & Methods). Results are shown both for all complex predictions (All) and for the high-confidence subset (HC). #, number of; nr, non-redundant.

| Prediction type | Mycoplasma to | | | | | |
| | Yeast | | Fly | | Human | |
| | All | HC | All | HC | All | HC |
|---|---|---|---|---|---|---|
| # query complexes | 174 | 67 | 174 | 67 | 174 | 67 |
| # predicted complexes (nr) | 11 | 3 | 9 | 1 | 6 | 1 |
| Total # complex components (nr) | 125 | 8 | 86 | 6 | 39 | 3 |
| Average # components per complex | 13.64 | 2.67 | 12.0 | 6 | 7.17 | 3 |
| Total # novel components (nr) | 86 | 1 | 68 | 4 | 30 | 0 |
| Average # novel components per complex | 9.36 | 0.33 | 9.56 | 4 | 5.33 | 0 |
| # entirely novel complexes (nr) | 0 | 0 | 0 | 0 | 1 | 0 |

# Figure 6

Mycoplasma complexes leading to significant predictions in yeast, fly and human.

Comparison of the different sets of mycoplasma complexes that led to significant predictions in yeast (yellow), fly (blue) and human (red) through complex to network alignment using NetAligner [36] (grey arrows). Overlaps between the different sets are shown as a Venn diagram. Complexes found in at least two species are listed.